Hendrik: Conversational agent for course evaluation

Pietro Camin
University of Twente
p.camin@student.utwente.nl

Armein Dul
University of Twente
a.m.dul@student.utwente.nl

Ioana Frincu
Univeristy of Twente
i.frincu@student.utwente.nl

Khiet Truong (Supervisor)
University of Twente
k.p.truong@utwente.nl

Ella Velner (Supervisor)
University of Twente
p.c.velner@utwente.nl

ABSTRACT

The University of Twente currently sends an online questionnaire to collect the opinions of its students regarding courses. The response rate is always low which hinders the improvements of courses due to lack of qualitative insights from students. This research aims to investigate if an Embodied Conversational Agent (ECA) can be used as a replacement for evaluating courses at the University of Twente. In this study, a Furhat ECA, named Hendrik, has two different conversational styles (formal and casual). These conditions were tested through a conversation with the students. The variations between conversational styles and the response strategy were confronted with a ground control group, all three conditions being later assessed and compared through the SASSI questionnaire. To validate if the responses can be regarded as qualitative and useful we relied on the teachers of the courses to compare them using the SSA method. The results showed how students find the formal ECA as the most appropriate and easy to use method for course evaluations, as it engages them the most while being easy to understand. However, answers were found most useful by professors when students interacted with the casual ECA, suggesting that further research should aim at building a better system capable of having natural conversations. Insights from this study on conversational styles for an ECA can be used as a baseline in building better performing agents, which is a real need for service oriented industries where collecting qualitative feedback is essential.

Author Keywords

Embodied Conversational Agent, course evaluation, conversational style, user engagement

INTRODUCTION

At the University of Twente, Bachelor's modules and Master's courses are evaluated by a so-called Student Experience Questionnaire (SEQ) [1]. Through this questionnaire, students are able to provide feedback on the quality of their education. The results show an overview of the opinion of the students towards a course or a module [1]. One of the advantages of online questionnaires is that it is possible to reach a large number of students and analyse the results easily [2]. However, the questionnaire can lead to less in-depth answers, as there is no interviewer available to, for instance, ask for clarification [2]. The study of Kim, Lee and Gweon [2] evaluated if adding interactivity reduces satisficing behaviour. Satisficing behaviour entails that respondents often engage less cognitively by choosing an appropriate, satisfying response instead of an accurate one [3]. In the context of this study, satisficing behaviour of students may occur due to the social desirability and respondent burden [4, 3], when filling in course evaluations. Students might show reluctance in expressing their true opinion, as they do not understand what happens to their feedback and why it is relevant to the faculty, besides having to respond to multiple long similar surveys for each course. Another issue of surveys is careless responding (CR) where respondents fail to read the instruction and incorrectly attend to an item [5]. CR is the result of the lack on intrinsic motivation, where respondents do not desire to respond accurately and feel like they have no responsibility to accurately fill in the survey.

This leads to several cases where SEQ might prove to be inefficient, most of them stemming from lack of intrinsic motivation, satisficing behaviour, and careless responding. The main external factor, which can be partially influenced, is the engagement of the respondent. The focus of this study is to tackle the careless responding by engaging the students into giving the course evaluation.

Hence, an alternative approach is being researched to engage the students and evaluate the courses. The alternative approach is to let the students interact with an Embodied Conversational Agent (ECA). The main difference is having a conversation to evaluate a course, rather than filling out a questionnaire. To engage the students, it is of importance to focus on the conversational aspect through discussing topics in depth and creating an optimal experience where the student can freely express and engage without any hold backs. Still keeping in mind that generating these answers should be useful for the teachers at the University of Twente. In the context of this study, these answers are desired to be qualitative. Qualitative answers can be described as the answers that normally are not discovered with quantitative research. This general description has been derived from the study of [6], where they have researched quantitative and qualitative feedback techniques. A conversation with an ECA can be seen as a qualitative feedback technique.

Consequently, it is currently unclear how the agent should behave to engage the users in providing the expected and desired qualitative answers. Therefore, the research questions can be formulated as follows:

Research Question: How does the conversational style of an embodied conversational agent (ECA) affect the quality of the responses from University of Twente students towards the course evaluation, when compared with the existing web-based questionnaire (SEQ)?

The overall aim of this research was to design an ECA, which was named Hendrik, be able to engage users in interactive conversations to gather qualitative answers for a Master's course evaluation. In the upcoming section, the related work with respect to conversations with an ECA and conversational styles of an ECA are described. Followed by the method that was used for the alternative course evaluation with an ECA. Thereafter, the results of the study are presented and elaborated in the Discussion section. At last, the conclusion of this research is given, having in mind the above-mentioned research question.

RELATED WORK

The main goal of this research is to alter the experience of evaluating academic courses by transforming the questionnaire into a conversation. Moroney and Cameron [7] explain that it is in the hands of questionnaire designers to shape the wording and personalise the questions to obtain more accurate and qualitative information from the respondents. In the next paragraph, there will be shown that previous studies focusing on the interaction element of a conversation within a survey, have highlighted two comparison elements: the conversational style and the channel (platform).

Conversations with an agent

According to Tannen [8][p.288] conversational style can be described as "... the use of specific linguistic devices, chosen by reference to broad operating principles or conversational strategies. The use of these devices is habitual and may be more or less automatic". Tannen's work [8] is an important pillar in defining the elements which construct the conversational style, those being divided into four main categories: (1) topic, (2) pace, (3) expressive para linguistic and (4) genre. For creating the conversation flow between the student and the agent, these characteristics can be considered as guidelines.

However, human to human conversation is seen vastly different compared to the human to conversational agent conversation. The latter is perceived as transitional, where people cannot and do not want to build a relationship with a "tool" [9]. Clark highlights in his study that human-agent interaction should be seen as a genre of conversation, where the context of the conversation and the utilitarian purpose of the agent are more important factors than fostering relations, which is an essential perspective to understand [9]. This will help in managing expectations form both researchers and participants regarding the conversation with Hendrik. To evaluate the effect of this perspective in the student's engagement within the interaction, two types of conversational styles will be used for this study: a casual and a formal style, inspired by Wambsganss's experiment [10].

The study conducted by Wambganss attempted to change a university course evaluation from text-based survey to a conversation with a chatbot [10]. Such research as Wambganss's served as a guideline in this research. However, instead of a chatbot, a human-like virtual conversational agent is used. The purpose of having a humanoid virtual agent instead of a text-based chatbot is to investigate if the visual embodiment combined with non-verbal behaviour (nod, gaze, facial expressions, etc.) can promote a fluid conversation between agent and respondent, while minimizing the social desirability which occurs in human-to-human conversations. So far, there is evidence that the visual component plays an important role in facilitating trust within the human-agent interaction [11]. For Schmutzler et al. [4], the embodiment of the agent had no significant impact on self-disclosure, the main fault being the non-interactive animation. The design of the conversational agent will take into account its ancestors' configurations and limitations, such as Rhea and Great, by placing priority on the conversation's characteristics where the embodiment serves as an appropriate support for it [12, 13, 14].

In this study, the Furhat virtual agent [15] will be used given its wide range of available configurations, especially for linguistics and non verbal behaviour, together with its extended customisation libraries. This allows for creating a virtual agent which can mitigate the limitations in human-agent interaction (uncanny valley, inflexible tonality, static non-verbal behaviour).

For evaluating and comparing the conversational styles between the agents, there will be made use of the alignment and speech markers used in Thomas et al. [16] who investigated the conversational style in voice agents based on Tannen's conversational characteristics [16] (see Appendix A).

In summary, the conversational style of the agent and the shaping of the questionnaire, including the questions as relevant conversational items, are essential aspects to achieve an increased engagement and enjoyment of respondents. In the upcoming section, the theoretical difference between the casual and formal conversational styles are explained in depth.

Formal conversational style

A formal conversational style resembling the way questions are shaped in most current questionnaires. Such a style was perceived as "stiff" and "cold" by participants in Wambsganss et al. [10]. Based on the context and topic discussed, participants in the study of Shamekhi et al. [17] confirmed they preferred a conversational agent which matches their conversational style. Thus, there are instance where the formal style is preferred such as official, informative and problem solving oriented scenarios where the formal conversational style can facilitate direct and convenient communication. However, in some cases the formal style is seen as impersonal and is less flexible in touching upon sensitive topics in a conversation due to the lack of common ground and absence of an equal relationship [9, 18].

To confirm literature's findings, an ECA with a formal conversational style will be designed showing minimal non-verbal behaviour micro-expressions (blinking, nodding, etc). The ECA will address the user in a personal manner while maintaining a similar structure as the SEQ, following a waterfall conversation. The formal style main goal is to identify difference in the overall quality of answers when they are spoke instead of written form. By directly talking to the agent, the conversation incorporates spontaneous interaction [19] and allows students to reflect upon their experiences and feelings during the course by talking out loud [20].

Casual conversational style

One of the relevant findings from Kim's et al. [2] study is that the casual conversational style of the agent encourages self-disclosure emulating a high social presence, leading to the respondents perceiving the whole survey experience in a positive light. The aim of the casual style is to rephrase the questions from formal to informal, representing a warmer and friendly attitude of the CA through voice, pauses, non-verbal behaviour and choice of words for asking more in-depth answers. Casual conversational style also includes non-verbal behaviour such as head titles, micro expressions and nodding at the right time which allows to increase the perception of "warmth" and active listening to build a rapport between the human and the agent [21].

Below, one of the questions is presented in the three cases, which are the focus of this study. The first question is reproduced from the general SEQ. The output is a Likert-scale answer. The second question is in the formal conversational style. At last, the third question is in the casual conversational style.

Current SEQ question: "The feedback during the course gave me sufficient information for further learning".

Formal style: " How would you rate the feedback given in this course? Do you feel it prepared you for further learning by informing you enough? "

Casual style: " Do you think that the feedback received during the course gave you enough information for further learning? You can think of any feedback, from teachers, peers or from other staff."

Qualitative answers

There is no standard definition of what a qualitative answer is. To define the meaning of more qualitative answers from the user, there has been settled upon a set of features to assess the quality, such as robustness of the answer, follow up answer match, correctness and accuracy with regards to the question and numbers of conversational topics found in an answer [22, 23]. To fairly compare the three results (the SEQ, Formal Hendrik and Casual Hendrik), we will use the expertise and knowledge of the teachers to assess the quality and usefulness of the answers, mainly for two reasons: 1) course evaluations are meant as improvement guidelines for course coordinators, and 2) effects enabled by the teachers based on the results are the main factors affecting student academic gain [24].

So far, one study firmly confirmed the efficiency of a conversation in course evaluation. Kim, Lee and Gweon's [2] text-based chatbot succeeded in keeping the user engaged and gathering more qualitative data in comparison with the typical web-based survey. Within the context of this study, the aim is to explore if indeed a human-agent interaction can improve the quality of answers given by students regarding university courses which they recently have followed.

METHOD

The section is composed of four parts. An in-depth description of the two ECA styles, the methods used to evaluate the systems and the conversations, the participants of the experiment, and the procedure of each part of the study.

ECA Styles

To design a conversation for an ECA to evaluate University of Twente courses, first, the generic SEQ has been received from the organisation that develops the questionnaires at the University of Twente. To create the ECA, the virtual version of the Furhat robot [1] is coded to serve as an embodied conversational agent. The agents are programmed using the Kotlin language, version 1.7.10-release-333. The version of the Furhat Software Development Kit (aka *SDK*) is *Generation2: Stable* (aka *Gen2: Stable*).

Formal

The formal system is built using a waterfall conversational flow based on a fixed list of questions. Each question is a reformulation of a SEQ statement. For example, the SEQ statement "*The course topics were relevant for the educational programme*" was reformulated as ""*Were the course topics relevant for the educational programme?*". The purpose of the second condition is to compare the difference between the written evaluation with a spoken one and monitor if the quality of answers is better, and if the participant engages and enjoys more talking to an ECA than filling the SEQ. Regarding the technical aspect, all topics present in the SEQ were made as States in Furhat where, after the response of the user – using the *OnResponse()* function – Hendrik jumps to the next state. From this point Hendrik asks the next question. What gives the name "formal" to the style is the ECA linguistic manner, which is similar to the one used by the University of Twente in official communications. After the ECA asks a question, it accepts the answers independently from their content, following a system of timers presented in Table 1. In total, besides the informed consent and the storage of conversation questions, Formal Hendrik has 15 questions regarding the course evaluation.

The *Jamie* face and *Brian* voice were used, both provided by the Furhat package. All the settings related to the ECA's expressions, such as MicroExpression behaviour, were inherited from the Furhat Advance template skill which can be found in the SDK when creating a new skill. The agent's non-verbal behaviour shows the default micro-expressions: minimal left

[1] https://furhatrobotics.com/

and right tilts of the head, eye movements and blinking, and facial movements like raising eyebrows. The mouth moves synchronously to what the ECA says.

Casual

The casual system discourse flow consists of a dynamic discourse structure, based on the questionnaire list of questions, but that can adapt to users' answers. Rather than going through the topic linearly, the topics are chosen based on the response of the participants and depending on the length of the answer combined with the sentiment analysis. After every time that the ECA asks a question, the ECA runs a cycle to decide if there is a need for further elaboration or if it can move to the next question. The logic behind the cycle can be seen in Figure 1.

What gives the name "casual" to the style is the ECA linguistic manner: it uses a language style similar to what a friend of the user could use. Before asking for elaboration or a new question, the ECA's output is modelled accordingly to the sentiment analysis of the user's last reply to make the agent appear friendly and compassionate. The sentiment analysis happens through the built-in natural language understanding of Furhat where based on hard-coded dictionary files, the program associates an utterance with an intent. The logic cycle that regulates the sentiment analysis can be seen in Figure 2. The length of answers is measured through the "it.speech.length" function and we chose initially 3 seconds based on literature, but later changed it to 7 seconds based on the conversation from Formal Hendrik. Casual Hendrik used the VADER sentiment lexicon [25] due to the generality of it and the sentiments were divided into positive and negative. For the elaboration on feedback, we used the *call(state)* function where we activated the state elaboration where Hendrik asked for a more explicit answer or to summarise it and then depending on the sentiment analysis it replies positively or empathically (for negative sentiment).

To regulate the conversation dynamics, the same timers used for the formal agent were integrated (see Table 1). To create the virtual persona, the *Jamie* face and *William* voice were used, both provided by the Furhat package. If the face remained the same as the casual version of the ECA, we decided to change the voice since it enables functions that emphasise some of the agent's spoken words while talking. However, we acknowledge that this choice introduces more variability in the experiment. As the formal agent, the casual ECA shows some default non-verbal micro-expressions. However, while using the same settings as the Advance Template skill, it also implements non-verbal expressions – like nods, big smiles, or surprised expressions – that trigger at specific points of the conversation. For example, the ECA sometimes nods right before moving to the next topic to feedback that it listened to what has been said. The mouth moves synchronously to what the ECA says. All code can be found on GitHub, in the links in the Appendix.

Timeout timers

Both the formal and casual ECAs have internal timers to help regulate their behaviour during the conversations.

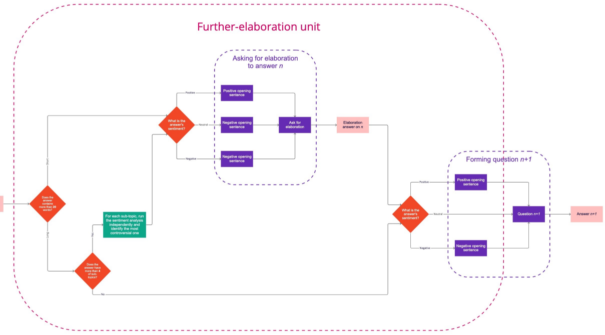

Further-elaboration unit
Asking for elaboration to answer *n*
Answer *n*
Does the answer contains more than 20 words?
Short
Long
Does the answer have more than 4 of sub-topics?
Yes
No
For each sub-topic, run the sentiment analysis independently and identify the most controversial one
What is the answer's sentiment?
Positive
Neutral
Negative
Positive opening sentence
Negative opening sentence
Negative opening sentence
Ask for elaboration
Elaboration answer on *n*
Forming question *n+1*
What is the answer's sentiment?
Positive
Neutral
Negative
Positive opening sentence
Negative opening sentence
Question *n+1*
Answer *n+1*

Figure 1. Flowchart of the cycle that occurs when the formal ECA asks a new question.

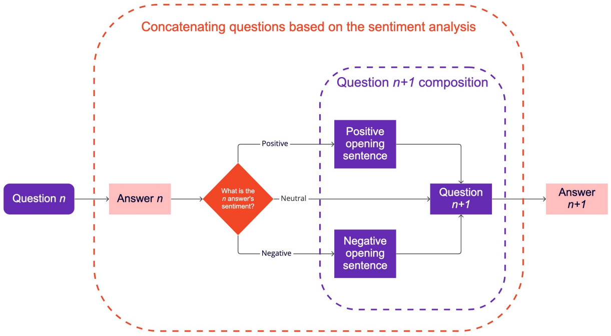

Concatenating questions based on the sentiment analysis
Question *n+1* composition
Question *n*
Answer *n*
What is the *n* answer's sentiment?
Positive
Neutral
Negative
Positive opening sentence
Negative opening sentence
Question *n+1*
Answer *n+1*

Figure 2. Flowchart of the sentiment analysis

The first timer limits the length of a user's answer to 59 seconds. Its timeout makes the agent move to the next stage, i.e. the next question for the formal ECA and the answer's elaboration process for the casual one.

The second timer regulates the time the Furhats wait before repeating a question. The timeout triggering the event happens after 6 seconds of silence while the ECA is in listening mode.

Finally, after 3.5 seconds of silence while in listening mode but with an answer already begun, the timer that controls the ECAs going to the next stage in response to a user's reply activates. Therefore, a user must wait 3.5 seconds after completing a reply before the ECA continues with the questionnaire by itself (see Appendix, Table 1).

Variable	Time
endStillTimeout	3.5 s
noSpeechTimeout	6 s
maxSpeechTimeout	59 s

Table 1. Timers regulating the agent's listening behaviour.

Evaluation methods

Two methodologies were used to evaluate the interactions with the ECAs: the SASSI and the SSA augmented questionnaires.

By doing so, the study makes use of both quantitative and qualitative methodologies [26] to compare and evaluate accurately the three methods for course evaluation.

SASSI questionnaire

The SASSI questionnaire (shown in Appendix C) is used for quantitative data collection and consists of a self-report measure of the UX of the evaluated system. It focuses on measuring the general speech-system usability as well as the user's experience. The choice for choosing SASSI was based on a comparison between multiple assessments tools for user experience (UX) in conversational interfaces [27], where SASSI fit the best with the context of this research due to measuring dimensions such as Enjoyment/Fun, and Frustration, but also as an overall combination of all other UX dimensions when interacting with a voice interface. While SASSI is created for speech-related technology assessment, it was slightly modified so that it can apply to the assessment of the SEQ method as well, thus evaluating the whole experience of filling in an online questionnaire and the system as a whole. The only words changed were where "speech" or "speak" was present in the statements, we added a "write".

In total, the SASSI consists of 6 categories about likeability, speed, habitability, cognitive demand of the user, system's response and habitability. Each statement had to be rated on a Likert scale from 1 to 6, where 1 means "Strongly disagree" and 6 is "Absolutely agree". An open-ended question was added at the end where the participant had the freedom to add any remarks/suggestions of improvement regarding the system, which helped in gathering the user's opinions on all of the three methods.

SSA augmented questionnaire

To measure the quality of a response given a context, the conversation between the agent and the students will be evaluated by the professor who coordinates the course. Inspired by the work of Adiwardana et al. – which used the sensibleness and specificity average (SSA) metric [28] calculated from the sensibleness and specificity scores – we tested three variables: sensibleness, specificity, and utility.

The first score, sensibleness, measures if an utterance fits the general context, evaluating whether a model's response is appropriate in light of the surrounding information and does not conflict with anything expressed until that moment. Humans frequently take this fundamental criterion for communication for granted. At the same time, generative models frequently struggle to meet this requirement, escalating in unintentionally rewarding models for playing it safe by consistently producing brief, generic, and uninteresting responses when sensibleness alone is utilised to judge models. The GenericBot algorithm [28], which always responds to queries with "I don't know" and to statements with "Ok," achieves a sensibleness score of 70%, even exceeding certain sophisticated dialogue models.

A response's specificity to a particular context is evaluated using the second score, specificity. A user would say, "I adore Florence," and the model might reply, "Me too." In this case, the model would receive a score of 0 for specificity because this response could be applied to a variety of situations. "Me too. I love Santa Maria del Fiore Cathedral," would receive a score of 1. According to Adiwardana et al., Meena closes the gap to the human performance average in the SSA metric [28]. However, when applying the SSA metric to questionnaire answers, sensibleness and specificity are insufficient to assess a dialogue's quality. For instance, a logical and pertinent solution to the question "What are the strong points of this course?" would be "Having a professor is a strong point for sure." The response "The professor could explain the topics well while keeping the class entertained" would be a deeper and more gratifying alternative. This is the reason for introducing the usability metric in the questionnaire, making it more similar to the Sensibleness, Specificity, Interestingness (SSI) metric [29] used by Thoppilan et al. where they introduced "Interestingness" as a third Boolean variable to measure whether an utterance is perceived as likely to "catch someone's attention" or "arouse their curiosity", or if it is unexpected, witty, or insightful [29].

For each of the three metrics, professors checked a box if an utterance was rated sensible, specific, or useful, resulting in three Boolean values.

While the score has been thought for evaluating language models for dialogue application, this study calculates it for participants' utterances.Sensibleness contributes to understanding whether the students' answers fit the context or if a question is misinterpreted.If an utterance is labelled as sensible, we further ask professors to determine whether it is specific to the given context or could be generally used for different situations. This helps us identify if the students' answers are too generic to produce valuable feedback.Usability's objective is to measure the quality of a student's response by asking professors if the answer provides good feedback for the course's improvement.

Four open-ended questions are found at the end of the questionnaire to assess the difference between the three conditions and to provide the opportunity to present eventual remarks, allowing the collection of qualitative insights. Such questions are the following:

- What kind of differences did you notice between the three methods?
- Could you describe if one or more of these methods gives you insights into your teaching and your course?
- Could you describe if one or more of these methods helps to improve your course?
- Are there any remarks that you would like to point out?

Participants

Participants of the study are divided into two groups: students and professors. Experimenting with the three conditions was done by the students; the lecturers compiled the SSA to rate the feedback received through the two ECAs conditions. The recruitment of the participants was done via personal recruitment.

Students

The study was between subjects, meaning participants experienced only one of each condition. In total, 20 participants were personally recruited via convenience sampling. The selection for which one of the three groups (SEQ, Formal Hendrik, Casual Hendrik) the participant should attend was made based on the availability of the participant. No other demographics of the participants, besides which study they are following and if the University of Twente is their primary enrolment university were collected in this study, to make sure that no identifiable information can be deducted about the participants. No gender quota or age range influences the participants' selection.

Professors

For the qualitative methodology, to test if the data obtained from the formal and casual conversations with Hendrik represent in-depth and qualitative answers which can help professors improve their courses, we asked the coordinators of the four courses evaluated by the students if they would be interested in rating the three condition's answers. All the professors agreed to take part in the study, analysing how the different systems and students interacted with one another by using the SSA augmented questionnaire. These professors weren't picked at random; rather, are those who taught the courses previously evaluated by the students. As a result, evaluations of the answers' quality in the filled questionnaire are more trustworthy.

Experimental design

The experimental study is designed to measure between subjects' experience for the three conditions (SEQ, formal Hendrik, and casual Hendrik). Inspired by the study of Wambsganss [10], where a formal and casual chatbot is introduced to students for course evaluation, our research also has a grounded comparison with the current method of evaluation, the SEQ. Therefore, this study attempts to answer the research question through two comparisons: 1) SEQ vs ECA and 2) which ECA is better in the context of course evaluation – the formal or the casual.

There are two factors used in this study for data triangulation [30], one being the three different course evaluation methods - one written as in the SEQ and two spoken as in the two versions of the ECA - and that each condition has unique participants per course, minimising the chances of a biased data collection where the student subconsciously compares the methods between each other.

In the experiment, characterised by three conditions, the participating students had to evaluate a course they attended during the third quartile of the 2021-2022 academic year. The courses were four, for each of which two students filled the SEQ, two interacted with the formal agent, and one talked with the casual agent – as shown in Figure 3. The data collection took place over three days, where a room on campus was booked for a duration of six hours each, accommodating a total of 8 participants per day (respectively, four for the last condition). First, the SEQ was evaluated; the next day followed the Formal Hendrik; five days later the Casual Hendrik was evaluated. Based on the results from the formal conversation, the casual Hendrik was adapted and tweaked to include more variety in the answers.

Participants were invited to fill in their preferred time slot of 30 minutes. Even though it was expected the experiments lasted shorter, it gave the researchers enough time in case an unexpected issue would appear, such as technical issues or the late arrival of the participant. Once the participant shows up to the indicated room, one of the researchers welcomes them, explains the purpose of the study, and the meaning of the participant's contribution, and orally reassures them that the participant's privacy is a priority. Then, the researcher hands them the information brochure and informed consent in duplicate.

As the focus is on the answers that the participants provide, the conversation between the ECA and the participant has been recorded. Through Furhat, the conversation log can be automatically stored as text in a JSON file. Furthermore, Furhat was configured to record the speech of the participants in WAV files to ensure the reliability of the data. To ensure the data is not lost, the conversation was also recorded via one of the laptops, using a voice recording application as a backup. The audio recordings were safely stored locally on the laptop SDD to be later deleted when conversations were transcribed with the help of the conversation logs. All the transcriptions were done within two weeks. To correct eventual errors in the ECA transformation from speech to text without losing the answer's meaning, the conversation transcripts were manually checked by the researchers and misspelt words were adjusted accordingly.

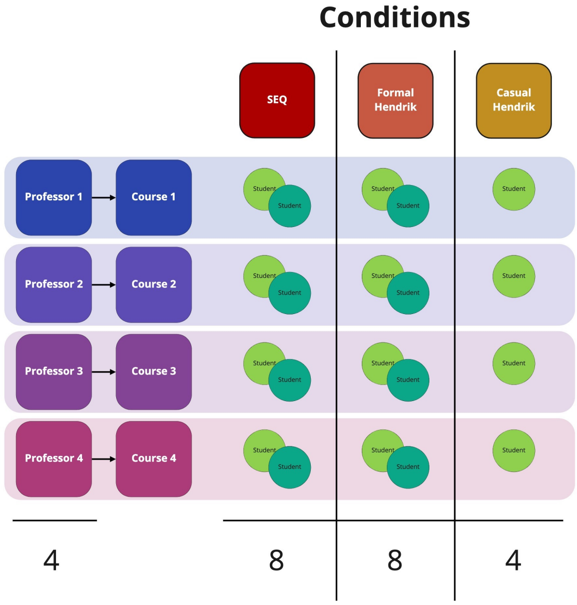


Figure 3. Scheme visualising the design of the experiment.

Condition 1: SEQ

As a control group, we asked 8 participants to fill in the standard SEQ from the University of Twente via the Qualtrics platform [2]. This first condition tested simulates the current course assessment done by the university. The questionnaire presented to the participants was identical to the sample one offered by the department which works on SEQ at the University of Twente. After submitting the mixed Likert scale and open-ended questionnaire, they filled out the SASSI questionnaire. To do so, participants were asked to recall the whole process of filling in the SEQ and evaluate the system based on their personal experience with it.

Condition 2: Formal Hendrik
To collect data for the second condition, participants were asked to interact with the ECA while the researchers left the room. The agent introduced itself, the experiment, and asked to agree to the privacy policy at first. Then, if the participant agreed, it asked questions relative to the course evaluation divided into five categories: (1) administrative, (2) about the course content, (3) about the teaching methods, (4) of general scope, and (5) if the participant had any suggestion towards the course improvement. After all the questions were answered, the agent asked the user if, considering what has been told, they wanted to submit the answers. When the interaction with the ECA was complete, participants proceeded with filling in the SASSI questionnaire to evaluate the agent UX. After the questionnaire, researchers could return to the room to thank the participants and give them a sweet participation gift.

No particular problems occurred while conducting the experiment. However, it once happened that the agent interpreted a negative response when asking for the privacy terms agreement. A possible explanation is that this occurrence was caused by the participant's accent and the inability of the agent to interpret it correctly. With the last participant, we thought that the same problem occurred too, but the laptop operative system deactivated the microphone autonomously instead. While not able to figure out why that happened, we reactivated the microphone and restarted the experiment smoothly.

Condition 3: Casual Hendrik
As previously mentioned, the Casual Hendrik was built and adjusted based on the participants' suggestion for formal, to differentiate between formal and casual styles. While this might impact some of the independent variables such as the intensity level of non-verbal behaviour and the formulation of questions, the modifications added during the experimental study helped to avoid similar feedback from participants regarding the ECA conversation as well as allowing us space to fix technical errors, e.g. the default listening limit of Furhat which interrupted a significant part of the participants. While testing Casual Hendrik, there occurred one particular problem. As the conversation of Casual Hendrik had been altered based on the conversations students had with Formal Hendrik, the state machine had been changed, i.e. the flow of the conversation. Therefore, Hendrik did not understand one of the participants and kept repeatedly asking a particular question. There was enough time between the data collection sessions thus the researchers were able to change the flow of conversation and solve the technical issues of Hendrik getting stuck on one question.

The SSA survey
After the data collected from the third experiment was cleaned and processed, to measure the quality of a response, the conversation between the agent and the students were evaluated by the professor which coordinates the course. In this way, we test the sensibleness, specificity, and utility of the participant's answers with our implementation of the SSA questionnaire.

The first two metrics derive from SSA, the Sensibleness and Specificity Average [28] while the latter aims at understanding how useful the piece of conversation is. In particular, sensibleness is measured by asking if an utterance fits the general context. It contributes to understanding whether the agent's questions feel natural in the discourse flow and if the answers of the students fit the context or if the question is misinterpreted. We further ask the professor to determine whether it is specific to the given context or could be generally used for different situations. This will help us identify if the framing of questions consistently elicits unproductive answers. Finally, usability will highlight if professors find the students' answers helpful. The metric will also measure the quality of a student's response, by asking if the answer provides good feedback and/or could be used to improve a course. For every metric, professors will have to evaluate utterances by answering a yes/no question, resulting in Boolean values.

Ethical considerations
To make sure this research is ethically viable, an ethical request has been documented and has been approved by the Ethical Committee of the faculty of Electrical Engineering Mathematics and Computer Science (EEMCS). Participants participated voluntarily in this study. The participants could withdraw before, during or immediately after the experiment. Moreover, there were no risks involved in participation. With respect to the system, having a conversation with the ECA does not lead to suggestive prompts. Instead, the ECA responds to the provided answers of the participants. At last, at all times, the identity of the teachers and the students, including their gathered data, who have participated in this study, will remain anonymous. The conversation between the participant and the agent has been audio recorded (with permission as stated in the consent form). Only participants who agreed to this, were included in the research. The audio recording was necessary, as the current Speech Technology, which is integrated into the Furhat, is not as reliable as an audio recording. It might happen that the Furhat misunderstood what a participant said. Therefore, the audio recording was used to transcribe the conversation. After transcribing the conversation of the participants solely the written format of their answers is shared with the teachers. The audio files have only been kept on the hard drive of one of the researchers for a maximum of two weeks. After this time had passed, the audio files were deleted.

RESULTS

To highlight the differences between the three experiments, the Results section accentuates the most relevant findings from the SASSI questionnaire through each of its six sections. The Likert scale of SASSI ranges from 1 (strongly disagree) to 6 (strongly agree).

[2] Qualitrics homepage: https://www.qualtrics.com/

System Response

For the SEQ, participants reported that the system is almost completely predictable (M = 1.88), but still dependable (M = 4.00) which would explain why the statement "*The system makes few errors*" scored the highest (M = 4.25). Most statements regarding the accuracy, efficiency, and reliability scored close to neutral (M = 3.2).

The Formal Hendrik scored quite similarly to the online questionnaire, except for the statement "*The system didn't always do what I wanted*" which was the highest rated (M = 4.00). Overall, the formal Hendrik was not perceived as having a better, more reliable response than the SEQ.

The Casual Hendrik also score overall between 2 and 3 on average. Participants did not perceive casual Hendrik as being more reliable or efficient in how it interacts, with one remarkable statement being "*The interaction with the system is efficient*" which scored the lowest out of the three conditions (M = 2.75). The main factor might be the expectations of participants regarding the system, where the system was seen as unpredictable and not acting according to what the participants expected (M = 3.75).

Likeability

The SEQ was perceived as easy (M = 4.00) and clear to use (M = 4.25). However, the SEQ scored low on friendliness (M = 2.00) and pleasantness (M = 2.13), and out of all the three conditions, it scored the lowest on enjoyment (M = 1.88).

Formal Hendrik scores are high on ease of use (M = 4.25) and on clarity (M = 4.25). Overall, the participants' scores were leaning toward positive values (all scored above a mean of 3, with a low standard deviation), showing that the system was perceived as pleasant and friendly compared to the SEQ. Formal Hendrik was perceived as more enjoyable than the other two conditions (M = 4.13).

For the Casual Hendrik, scores were neutral, with the exception of "*I was able to easily recover from errors*" where participants disagreed almost completely with the statement (M = 2.00). This might be because of the sentiment analysis conditions where the participants felt it was repetitive and that the ECA was expecting a specific answer from them. Casual Hendrik was not perceived more positively than the Formal Hendrik, scoring lower on friendliness (M = 3.00) and pleasantness (M = 3.00).

Cognitive demand

Overall, the scores between the three conditions were similar, with the exception that Formal Hendrik was ranked the best by being the easiest to understand on how use (M = 4.13). The main difference between the ECA (both conditions) and the SEQ is that while interacting with an agent requires more attention from the participants, it evokes more emotional responses from the participant (eg. calm, tense) compared to the SEQ. Still, this result is not outstanding to be considered significant yet due the large variance in participants' answers. If perception of the ECA and expectation of the conversation were pre and post evaluated, then we could have had better insights into the emotional response which directly impact the cognitive demand.

Annoyance

The SEQ was perceived as being the most boring interaction out of all three (M = 4.25), the less flexible (M = 4.00) and the most repetitive (M = 4.15).

Even though Formal Hendrik had the same questions in the same sequence as the SEQ, it was perceived as being the opposite of boring, less repetitive than the SEQ, and scored the highest on the system of use (M = 4.25). System of use mainly encompasses the usability and ease of use, thus participants found Formal Hendrik easier to interact with.

With an average score differential of 0.25 points less, Casual Hendrik performed similarly to the formal condition, with the only major exception being that it was perceived as more irritating (M = 3.15) than the Formal Hendrik (M = 2.15). The system's perception as repetitive and irritating is most likely due to the technical implementation of the elaboration system.

Habitability

Surprisingly, all three systems scored similarly in the habitability section, even for statements such as "*I sometimes wondered if I used the right word*" and "*I always knew what to write/say to the system*". Formal Hendrik proved to be the most consistent out of all three because participants could more easily keep track of the conversation (for the statement "*It is easy to loose track of where you are in an interaction with the system*", M = 2.13 compared to M = 3.00 for the SEQ and M = 3.1 for the casual condition).

Speed

Even if by less than a point, the SEQ was perceived as having the fastest interaction (M = 1.88) out of all three. Participants were neutral about the systems' speed (For "*The interaction with the system is fast*", all three conditions had an average of around 3.5).

Average completion time

It is worth mentioning the average completion time of each condition, which can be easily confronted in Table 2. For the SEQ questionnaire, participants spent 17.5 minutes on average. However, when eliminating the outlier (of 36 minutes, due to dyslexia), the average time spent on completing the questionnaire is 14.4 minutes. With the formal Hendrik, the average is 6.3 minutes; while 11.3 minutes is the average with the Casual Hendrik.

Condition	Average completion time
SEQ	17.5 s
Formal	6.3 s
Casual	11.3 s

Table 2. Average completion time of each of the three conditions.

Feedback

Most feedback on the SEQ was regarding its content. How "*The system often doesn't reflect the specific types of education in the questions. That is, a questionnaire like this never seems to evaluate whether you preferred lectures vs tutorials, online* *vs offline, project vs exam. That kind of stuff is really relevant, even though it is not explicitly asked about*" and "*Forms are typically too rigid and do not leave space for actual opinions*". About speed, one participant simply suggested to "*Make it quicker, and shorter.*" Thus, the general format and the closed-ended questions are weak points of the SEQ.

The feedback on Formal Hendrik mostly regards the technical implementation of the agent's timer for moving to the next question (3.5 seconds of silence), resulting in the participants' interruption and not providing enough time to formulate an answer after interjecting with sounds such as "uhm..." (e.g., "*It's quite enjoyable. Though maybe the interaction could wait for me to finish my sentences when I have to tell a little more, although I might also just lost track of myself a little*," "*Wait longer to go to next question: it should understand that a 'uhm...' sound means that I am still thinking about the answer.*" Interesting remarks from participants also highlighted the scarce non-verbal feedback of the system as well as the missing repetition of the question when users asked for it. One participant stated: "*It felt more like a questionnaire read out loud, which feels weird. To me, it feels like the system is meant to have a more natural conversation, but it poses some pretty unnatural questions for that conversation. I think it can be improved by using different questions and perhaps asking multiple follow-up questions within a topic*" which is an interesting remark on the purpose of the ECA in the context of course evaluation and if in such a scenario a natural conversation is needed or not.

Issues mentioned by participants regarding Casual Hendrik were connected to the elaboration function and the knowledge graph of the ECA. "*It felt like the system needed a specific word to continue. However, in an answer these words might not always be recognisable or even present, so when an answer is rejected and requires something more specific use the words that need to be in the answer when asking for better answers*" or " *It would help if the system knows what the course is about (machine learning), so if a person points out something, the agent knows what the student is talking about.*"

SSA results

The recruited professors of the four chosen courses filled in the SSA questionnaire to validate student participants' answers. These sample sentences from the SSA can be found in Appendix D.

First, when examining each course separately, it becomes clear that the ways in which the professors have assessed the participants' responses vary. This could be because the professors did not interpret the questionnaire in the way that it was intended. As Adiwardana et al. [28] point out, the sensibility metric has been included as an extension to the specificity one, meaning that a sensible answer implies it is also specific. However, in the case of this research, there is a chance that the professors have understood the SSA questionnaire in a different way, making it problematic to state general conclusions on both sensibility and specificity.

Still, this research includes the usability metric, which proves to be a valuable addition (as shown in Table 3). Except for the professor from course 1, there is a positive trend in usability looking at the averages from the SEQ, the Formal condition and at last the Casual condition. For instance, the professor of course 2 thinks the utterances of the participants through the SEQ are 0.06 useful, the utterances through the Formal condition are 0.15 useful and the utterances through the Casual condition are 0.74 useful.

This positive trend is also noticeable in the total averages per course evaluation condition, which can be seen in Table 4. Between the SEQ and the Formal condition, there is a small increase (from 0.24 to 0.28). The average usefulness of the Casual Hendrik is 0.68. This is a 183% increase with respect to the SEQ.

	SEQ			**Formal**			**Casual**		
	Sensible	Specific	Useful	Sensible	Specific	Useful	Sensible	Specific	Useful
Course 1	1.00	0.72	0.28	1.00	0.54	0.04	0.95	0.65	0.45
Course 2	0.31	0.38	0.06	0.35	0.58	0.15	0.17	0.35	0.74
Course 3	0.00	0.83	0.11	0.42	0.27	0.42	0.47	0.21	0.89
Course 4	0.56	0.44	0.50	0.15	0.38	0.50	0.32	0.42	0.63

Table 3. SSA and usability averages with respect to the four participated professors of courses at the University and with respect to the evaluation condition

	Average per metric		
	Sensible	*Specific*	*Useful*
SEQ	0.470588	0.602941	0.235294
Formal	0.480769	0.442308	0.278846
Casual	0.469136	0.407407	0.679012

Table 4. Averages of the SSA metric per evaluation condition.

Feedback

Next to the SSA and usability metric, the professors have been asked open questions about the utterances of the participants. One of them is about the differences between the three conditions. One of the professors points out that "*The casual mode had the possibility of adapting and re-framing the question based on the previous answer.*" Moreover, another professor states that "*I usually do the least with the SEQ score metrics. I always hope to find some concrete suggestions in the formal questions (yes/no answers do not help at all), so I saw at some points the agent leading to better insights (or, actually getting to a relevant point) where the formal method would not get to.*"

Considering the insights the conditions gave the professors, one of the professors points out that "*...generic complaints (which are ALWAYS, with every course on quality/amount/timing of Feedback, assignments or projects being unclear, etc..). the agent could oblige me by an immediate response if a student utters that the scope of a session was unclear ('did you read the reader, did you read the page on canvas prior to the class, did you read the description on OSIRIS below the first paragraph).*"

This results in the question of whether professors have an opinion on whether one of the conditions helps them improve their course. Three out of four professors answered this question and all agreed unanimously: Casual Hendrik. Still taken into account the comment "*I think the agent-based adaptive questionnaire could be trained to get more helpful insights, yes.*"

DISCUSSION

First, it is clear from the results that an ECA stimulates students, engaging them more than the SEQ questionnaire when talking about their experience during the course. Formal Hendrik scored the highest in the SASSI questionnaire, being faster to complete than the SEQ, and rated as more enjoyable and easier to use than the online questionnaire. An interesting finding based on the SASSI results is that the cognitive demand and habitability scored similarly in all three conditions, which means that the enjoyment of the user is a considerable factor in influencing engagement. Since speaking and speech were not perceived as one is more demanding than the other cognitively and participants worried the same about their choice of word in the answers, the only relevant variable in the engagement of the use remains the enjoyment.

A remarkable example of what happened during the conversation with the formal Hendrik is that a participant explained why they marked the course low: the concepts taught were repetitive because already learned in other courses. Thus, instead of relying on an average number, the course coordinators can have easier access to students' opinions, and students themselves have the opportunity to reflect on their experience while talking to the ECA. If on one hand, the current implementation of the ECA does not guarantee that a student will include what topics they already knew and in which previous course they learned them, on the other hand, the SEQ does not ask for such feedback at all. Therefore, by providing a conversation with a more open structure, students have the chance to include other types of suggestions.

Up until this point, part of our research question has been answered. However, the casual conversational style with functions, such as elaboration on sentiment analysis, that has been implemented in this study, has not been proven to be a better option than the formal conversational style of Hendrik. Even more, it was perceived as more annoying and repetitive. An explanation for this occurrence is that participants thought the ECA expected a specific answer. The main reason may lie in the interpretation of the elaboration answers, which in the context of course evaluation did not put the participant at ease. It only further increased expectations for a human-to-human conversation with the ECA; an expectation that was obviously not met. In comparison with its' ancestor, ELIZA [31], the context and the complexity of the Casual Hendrik were perceived as underwhelming by participants.

Another pertinent observation made by one of the participants was that, since they were speaking to the agent rather than writing answers down, they were being more direct and critical. The participant described how they did not have to carefully rephrase utterances during the evaluations, and expressed their true opinion without many filters. This has been a positive effect of what has been found in the literature by [3], i.e. the opposite of satisficing behaviour.

However, a smaller group of participants had more negative experiences with the agent, finding it unnatural to have a conversation with the ECA. One of the participants indicated to have a preference towards the normal course evaluation method (the SEQ), as it mostly has Likert scales preventing the need of justifying an opinion. Furthermore, both Formal and Casual Hendrik lacked some grounding confirming that participants were understood – an interesting point that highlights the expectation of humanisation of the agent. Another problem that occurred with the Formal Hendrik was that some participants were interrupted regularly by the system (due to a timeout of the maxSpeechTimeout timer – see Table 1), even when still talking. Also, the time delay when the user ended his utterance and Hendrik started a new question (regulated by the endStillTimeout timer) was sometimes perceived as too long. This happened mainly for yes/no questions, suggesting that for syllabic answers a shorter timer should be used. Moreover, if the user took more than 6 seconds to think (see noSpeechTimeout in 1), Formal Hendrik prompted that he did not understand what the user was saying even if the user remained silent, repeating the question. Finally, in the case of Casual Hendrik, when the agent asked for the answer's elaboration to have more qualitative data, participants perceived them as if the ECA did not understand them.

With regards to the SSA, there already has been stated that it this validation metric has not been resulted into the results, where initially has hoped for. In this research, the teachers have filled out an online SSA questionnaire, instead of filling out the SSA questionnaire in a laboratory setting (similar to the condition of the participants of evaluating the SEQ, Formal and Casual condition). This means that currently there can't be concluded if the utterances of either the SEQ or the Formal condition, or the Formal condition are more or less sensible, and perhaps specific.

Based on the feedback of the teachers, an extra check on the knowledge, the materials that has been read, etcetera could be valuable to truly understand what parts of the course can be improved. As the teachers point out, there is always general complaints, but the comments that point towards specific unclarities can help the teachers.

LIMITATIONS

One of the biggest limitations of this study was altering multiple variables at once between conditions, such as how questions are formulated or a casual style with more expressive nonverbal behaviours than the formal one. Another limitation is the lack of qualitative data, which could explain the gaps in the SASSI scores. Moreover, due to time limitations, there has been an unequal division of participants between conditions which contributes to skewing the results. Furthermore, evaluating only 4 courses proved to be insufficient in understanding the teachers' perspective fully, and the fact that the SSA questionnaires were filled out online, prevented us from assessing their understanding of the three variables. The individual difference between participants might have also played a role in the conversation with the ECA expectations, appearing that with the Casual Hendrik participants expected even more interaction and a natural conversational flow, which might not have been met from their perspective.

FURTHER RESEARCH

If on the one hand, the experiment provided interesting insights on using an ECA for course evaluation, on the other hand, the results were not as clear cut as hoped. Further research can solve this problem by having an evaluation of the systems more qualitative than quantitative. It is advised to focus more on the qualitative methods in assessing the systems, otherwise, a lot of gaps such as the impact of the different nonverbal behaviour will remain uncovered. Further research should investigate more narrow research questions, where only one variable at a time will be changed between ECAs. Additionally, a clearer definition of the formal and casual conversational styles used by ECAs would be extremely beneficial in enhancing our comprehension of the situations and settings in which each conversational style is appropriate.

Further research should also thoroughly explore how both expectations and context influence human behaviour during an interaction with an ECA. Our study adjacently showed that these two factors impacted users' enjoyment, given that some of the participants expected Hendrik to be able of a human-like conversation while discussing the course evaluation in a formal manner.

Finally, new versions of the Hendrik will require better implementation of adaptability to the users' answers. One way of doing it is through the prediction of the answers' intent to ask for further information only when it is inevitable that answers will provide more useful hints for professors.

CONCLUSION

This research aims at investigating opportunities for using an ECA for university course evaluations. By testing two different conversational styles, along with the now implemented base ground, a step forward in improving the evaluation of university courses has been achieved.

The SEQ, which is currently the adopted solution, was rated by students as the least enjoyable, i.e. the least engaging to evaluate courses. This brings aversion towards its optional completion, lowering students' feedback rate. The formal ECA was perceived as more engaging than an ECA with a casual conversational style, contrary to our hypothesis and the literature. We think that the result is due to the implementation of the latter, which is more complex to model while also raising users' expectations.

The natural evolution of this work could take two different directions: either listening to teachers by improving the Casual Hendrik, based on the enhanced SSA useful score results; or listening to students and improving the more enjoyable formal condition by modifying the phrasing of the questions to extract better usable feedback that can help professors, based on SASSI results.

The formal condition suits the current situation the best. However, being the goal of the study, as stated in the research question, to improve the quality of the course evaluation answers –and, therefore, their usability–; future work should enhance the casual condition. Major improvements should involve refining its communication model so that students can feel understood during the conversation with the agent.

REFERENCES

[1] U. Twente, “EvaSys,” https://www.utwente.nl/en/educational-systems/about-the-applications/icto/evasys/, 2022, accessed: 2022-6-19.

[2] S. Kim, J. Lee, and G. Gweon, “Comparing data from chatbot and web surveys,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, May 2019.

[3] T. D.-L. Guin, R. Baker, J. Mechling, and E. Ruyle, “Myths and realities of respondent engagement in online surveys,” *Int. J. Mark. Res.*, vol. 54, no. 5, pp. 613–633, Sep. 2012.

[4] R. M. Schuetzler, J. S. Giboney, G. M. Grimes, and J. F. Nunamaker, Jr, “The influence of conversational agent embodiment and conversational relevance on socially desirable responding,” *Decis. Support Syst.*, vol. 114, pp. 94–102, Oct. 2018.

[5] M. Ward and A. W. Meade, “Applying social psychology to prevent careless responding during online surveys,” *Applied Psychology*, vol. 67, no. 2, pp. 231–263, 2018.

[6] D. C. Steyn, Carly and A. Sambo, “Eliciting student feedback for course development: the application of a qualitative course evaluation tool among business research students,” *Assessment Evaluation in Higher Education*, vol. 44, no. 1, pp. 11–24, 2019.

[7] W. F. Moroney and J. Cameron, “The questionnaire as conversation,” *Ergon. Des.*, vol. 24, no. 2, pp. 10–15, Apr. 2016.

[8] D. Tannen, *Conversational style: Analyzing talk among friends*. Cary, NC: Oxford University Press, Jan. 2005.

[9] L. Clark, N. Pantidi, O. Cooney, P. Doyle, D. Garaialde, J. Edwards, B. Spillane, E. Gilmartin, C. Murad, C. Munteanu *et al.*, “What makes a good conversation? challenges in designing truly conversational agents,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[10] T. Wambsganss, R. Winkler, M. Söllner, and J. M. Leimeister, “A conversational agent to improve response quality in course evaluations,” in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2020.

[11] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André, ““let me explain!”: exploring the potential of virtual agents in explainable ai interaction design,” *Journal on Multimodal User Interfaces*, vol. 15, no. 2, pp. 87–98, 2021.

[12] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan, “Embodiment in conversational interfaces: Rea,” in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 1999, pp. 520–527.

[13] J. Cassell, “Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents,” *Embodied conversational agents*, vol. 1, 2000.

[14] I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, and B. De Carolis, “Greta. a believable embodied conversational agent,” in *Multimodal Intelligent Information Presentation*. Dordrecht: Springer Netherlands, 2005, pp. 3–25.

[15] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, “Furhat: a back-projected human-like robot head for multiparty human-machine interaction,” pp. 114–130, 2012.

[16] P. Thomas, M. Czerwinski, D. McDuff, N. Craswell, and G. Mark, “Style and alignment in Information-Seeking conversation,” in *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - CHIIR ’18*. New York, New York, USA: ACM Press, 2018.

[17] A. Shamekhi, M. Czerwinski, G. Mark, M. Novotny, and G. A. Bennett, “An exploratory study toward the preferred conversational style for compatible virtual agents,” in *Intelligent Virtual Agents*, ser. Lecture notes in computer science. Cham: Springer International Publishing, 2016, pp. 40–50.

[18] R. E. Mayer, S. Fennell, L. Farmer, and J. Campbell, “A personalization effect in multimedia learning: Students learn better when words are in conversational style rather than formal style.” *Journal of educational psychology*, vol. 96, no. 2, p. 389, 2004.

[19] T. Iizuka and H. Mori, “How does a spontaneously speaking conversational agent affect user behavior?” 2022. [Online]. Available: https://arxiv.org/abs/2205.00755

[20] N. Novielli, F. de Rosis, and I. Mazzotta, “User attitude towards an embodied conversational agent: Effects of the interaction mode,” *Journal of Pragmatics*, vol. 42, no. 9, pp. 2385–2397, 2010, how people talk to Robots and Computers. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378216609003324

[21] T. Bickmore and J. Cassell, “Small talk and conversational storytelling in embodied conversational interface agents,” pp. 87–92, 1999.

[22] R. Mousavi, T. S. Raghu, and K. Frey, “Harnessing artificial intelligence to improve the quality of answers in online question-answering health forums,” *J. Manag. Inf. Syst.*, vol. 37, no. 4, pp. 1073–1098, Oct. 2020.

[23] J. Jeon, W. B. Croft, J. H. Lee, and S. Park, “A framework to predict the quality of answers with non-textual features,” in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’06*. New York, New York, USA: ACM Press, 2006.

[24] W. L. Sanders, S. P. Wright, and S. P. Horn, “Teacher and classroom context effects on student achievement:

Implications for teacher evaluation," *Journal of personnel evaluation in education*, vol. 11, no. 1, pp. 57–67, 1997.

[25] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text." [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[26] J. W. Anastas, "Quality in qualitative evaluation: Issues and possible answers," *Research on Social Work Practice*, vol. 14, no. 1, pp. 57–65, 2004.

[27] A. B. Kocabalil, L. Laranjo, and E. Coiera, "Measuring user experience in conversational interfaces: A comparison of six questionnaires," in *HCI 2018*. BCS Learning & Development, Jul. 2018.

[28] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like open-domain chatbot," Jan. 2020.

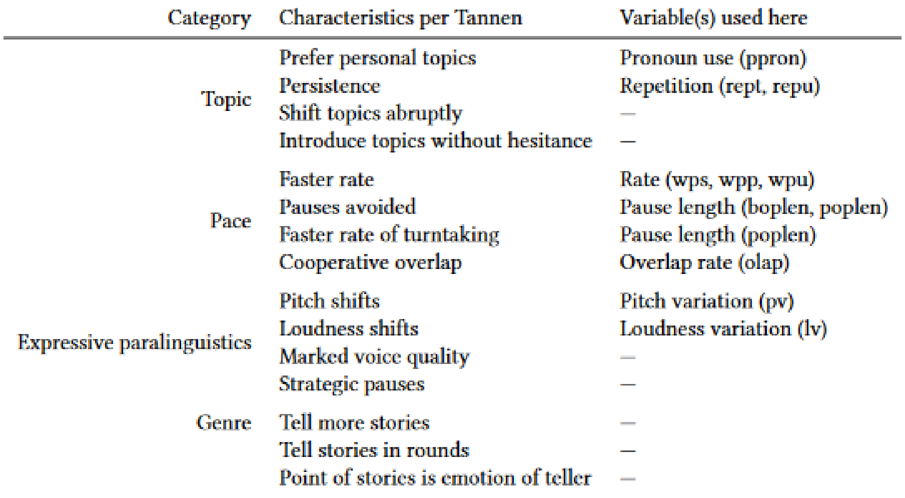
[29] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, and Q. Le, "Lamda: Language models for dialog applications," 2022. [Online]. Available: https://arxiv.org/abs/2201.08239

[30] J. Corner, "In search of more complete answers to research questions. quantitative versus qualitative research methods: is there a way forward?" *Journal of Advanced Nursing*, vol. 16, no. 6, pp. 718–727, 1991.

[31] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

APPENDIX

APPENDIX A: TANNENS'S CONVERSATIONAL STYLES CHARACTERISTICS



Category	Characteristics per Tannen	Variable(s) used here
Topic	Prefer personal topics	Pronoun use (ppron)
	Persistence	Repetition (rept, repu)
	Shift topics abruptly	—
	Introduce topics without hesitance	—
Pace	Faster rate	Rate (wps, wpp, wpu)
	Pauses avoided	Pause length (boplen, poplen)
	Faster rate of turntaking	Pause length (poplen)
	Cooperative overlap	Overlap rate (olap)
Expressive paralinguistics	Pitch shifts	Pitch variation (pv)
	Loudness shifts	Loudness variation (lv)
	Marked voice quality	—
	Strategic pauses	—
Genre	Tell more stories	—
	Tell stories in rounds	—
	Point of stories is emotion of teller	—

Figure 4. Characteristics that Tannen used for conversational styles [8]

APPENDIX B: SEQ QUESTIONNAIRE

1. Administrative
   (a) Which master programme do you follow?
   (b) At which university are you primary enrolled in (hoofdinschrijving)?
2. Content (1 to 5 Likert scale)
   (a) The learning goals and the related assessment criteria were clear to me.
   (b) The course topics were relevant for the educational programme.
3. Teaching (1 to 5 Likert scale)
   (a) The teaching activities challenged me to study.
   (b) The teaching staff encouraged me to think for myself.
   (c) I felt that the teacher had a good insight into how the students kept up with the content matter and acted adequately when necessary.
   (d) feedback during the course gave me sufficient information for further learning.
4. General
   (a) The knowledge and the skills gained in this course will not quickly fade away (i.e., they are lasting).
   (b) In general, the amount of study time I had to put into this course compared with the number of ECs granted was... Too few - Too much
   (c) What marks out of ten would you give this course? 1-10
   (d) Comments
      i. What is your opinion on the study material? (consider the book, lecture notes, guidelines, study guide, articles, etc.)
5. Your suggestions
   (a) What are the strong points of this course?
   (b) Do you have any suggestions to improve this course? Please elaborate.

Figure 5. Flowchart of the formal conversational style

APPENDIX C: SASSI QUESTIONNAIRE

System response

- The system is accurate.
- The system is unreadable.
- The interaction with the system is unpredictable.
- The system didn't always do what I wanted.
- The system didn't always do what I expected.
- The system is dependable.
- The system makes few error.
- The interaction system is efficient.

Likeability

- The system is useful.
- The system is pleasant.
- The system is friendly.
- I was able to recover easily from the errors.
- I enjoyed using the system.
- It is clear how to speak with the system.
- It is easy to learn how to use the system.
- I felt in control of the interaction with the system.

Cognitive demand

- I felt confident using the system.
- I felt tense using the system.
- I felt calm using the system.
- A high level of concentration is required when using the system.
- The system is easy to use.

Annoyance

- The system is easy to use.
- The interaction with the system is boring.
- The interaction with the system is repetitive.
- The interaction with the system is irritating.
- The interaction with the system is frustrating.
- The system is too inflexible,

Habitability

- I sometimes wondered if I used the right word.
- I always knew what to write/say to the system.
- I was not really sure what the system was doing/ what the goal was.
- It is easy to lose track of where you are in an interaction with the system.

Speed

- The interaction with the system is fast.
- The system responds too slowly.

APPENDIX D: SSA SAMPLE SENTENCES

SEQ

	Agent	Participant
Course 1	The feedback during the course gave me sufficient information for further learning.	3 (neither disagree or agree) [Sensible: 1, Specific: 1, Useful: 0]
Course 3	The feedback during the course gave me sufficient information for further learning	4 (Agree) [Sensible: 0, Specific: 1, Useful: 0]

Formal

	Agent	Participant
Course 1	Do you have any suggestions to improve this course, please elaborate?	yes I think there should be more reasoning behind the focus on reasoning behind the decisions and just more expectations from the teachers for example testing [Sensible: 1, Specific: 1, Useful: 1]
Course 3	Were the learning goals and the related assessment criteria clear?	Yes, the rubrics were announced, there were hand outs, euhm yes. [Sensible: 0, Specific: 0, Useful: 1]

Casual

	Agent	Participant
Course 1	Let me ask you something. During the course there might have been a moment, that you received feedback on assignments, exams, presentations etcetera. How did the feedback during the course help you for further learning?	I was able to reflect more on what I did and then gather better inside. [Sensible: 1, Specific: 1, Useful: 1]
	Wow, interesting answer! I missed the main conclusion, could you repeat again?	the insight, the feedback helped me for getting proper insights and reflecting on what I learned and how I can use it in the professional career [Sensible: 1, Specific: 1, Useful: 1]
Course 3	Let me ask you something. During the course there might have been a moment, that you received feedback on assignments, exams, presentations etcetera. How did the feedback during the course help you for further learning?	It didn't feel like it was helping much since it was not very clear since it was not very similar every time so it didn't feel like improving the parts that really helped. [Sensible: 0, Specific: 0, Useful: 1]
	Could you maybe try to be more specific so I can know how to categorize the answer?	bad because it didn't seem like changing the suggestions helped for the next week [Sensible: 0, Specific: 0, Useful: 0]

Table 5. Sample sentences of the SSA questionnaire and their results. For the Formal Condition, two different sentences are chosen, due to participants not answering this question when interacting with Formal Hendrik.